ON JOINT MAXIMUM-LIKELIHOOD ESTIMATION OF PCR EFFICIENCY AND INITIAL AMOUNT OF TARGET

H. Vikalo, B. Hassibi, and A. Hassibi

California Institute of Technology, Pasadena, CA

ABSTRACT

We consider the problem of estimating unknown parameters of the real-time polymerase chain reaction (RT-PCR) from noisy observations. The joint ML estimator of the RT-PCR efficiency and the initial number of DNA target molecules is derived. The mean-square error performance of the estimator is studied via simulations. The simulation results indicate that the proposed estimator significantly outperforms a competing technique.

1. SUMMARY

The polymerase chain reaction (PCR) is an *in vitro* technique for enzymatic replication of DNA fragments [1]. Applications of PCR [2] include genotyping, detection of infectious and hereditary diseases, genetic fingerprinting, etc. Typically, a given sample contains only a small amount of the target which needs to be detected and quantified. PCR replicates the target using two primers which serve as initiation sites for a DNA polymerase enzyme. The reaction is carried out in a buffer containing nucleotides used by the DNA polymerase enzyme to replicate the template. PCR amplifies the target DNA through a series of temperature-regulated cycles. A cycle consists of three distinct steps: denaturing, annealing, and extension. During denaturing, the sample is heated to break the hydrogen bonds between strands of target molecules, creating single-stranded fragments. Then, during annealing, the sample is cooled to the temperature at which primers will likely hybridize to the templates. Finally, in the last phase, the sample is heated to the temperature which is optimal for the DNA enzyme activity so that the primers are extended at an optimal rate. Ideally, at the end of the extension phase, there are twice as many double-stranded target molecules as there were at the beginning of the cycle.

Theoretically, the number of target DNA molecules doubles during each temperature cycle, resulting in an exponential growth. However, practical issues affect the replication process adversely and the efficiency of PCR – defined as the probability of generating a replica of each template molecule – is smaller than desired. Random nature of the underlying biochemical process leads to variations in the PCR yield. Moreover, creation of non-specific byproducts in the replication process further diminishes purity of the PCR product.

Probabilistic nature of the replication process is addressed in [3]-[6], where various stochastic models have been proposed. In [7], the mutations-related effects that plague the efficiency of PCR have been studied. The ultimate goal of PCR is estimation of the initial number of target molecules. A common approach is that of finding an estimate of the efficiency first, from which the initial number of targets is deduced next. In this paper, we find the joint maximum-likelihood estimate of the PCR efficiency and the initial number of target molecules.

Let x_0 denote the initial number of target molecules which we want to estimate. We assume that the efficiency of replication during both the background phase and the exponential phase is constant, and denote it by p. Furthermore, denote the number of target molecules at the end of the n^{th} cycle by x_n, and note that

$$x_n = (1+p)x_{n-1} + \tilde{x}_n, \tag{1}$$

where $\tilde{x}_n$ denotes the variations in the number of amplified molecules in the n^{th} cycle, and is a random variable with zero mean and variance $p(1-p)x_{n-1}$. It can be shown (see, e.g., [3]) that the mean of x_n in (1) is given by

$$E\{x_n\} = (1+p)^n x_0. \tag{2}$$

Furthermore, its variance can be found as

$$\sigma_n^2 = \frac{1-p}{1+p}\left[(1+p)^{2n} - (1+p)^n\right] x_0. \tag{3}$$

Imperfect instrumentation and other biochemistry independent sources create a noise which corrupts the measurements of x_n. We assume that the noise is additive Gaussian $\mathcal{N}(0, \sigma_{\mathbf{w}}^2)$, and denote it by w_n. Hence, the quantity measured is given by $z_n = x_n + w_n$.

Let us denote the number of temperature cycles in the background phase of RT-PCR by k. Therefore, the first measurement taken beyond the background noise level is z_{k+1}. Furthermore, denote the number of temperature cycles in the exponential phase by l. Hence, the last measurement taken before the efficiency starts rapidly deteri-

orating is z_{k+l}. Introduce a new variable, $\mathbf{y}$, defined as

$$\mathbf{y} = \begin{bmatrix} \frac{z_{k+1}-(1+p)^{k+1}x_0}{\sigma_{k+1}} \\ \frac{z_{k+2}-(1+p)^{k+2}x_0}{\sigma_{k+2}} \\ \vdots \\ \frac{z_{k+l}-(1+p)^{k+l}x_0}{\sigma_{k+l}} \end{bmatrix}.$$

Since $\mathbf{y}$ can be represented as a sum of large number of independent, identically distributed (iid) random variables, we invoke the central limit theorem to argue that the distribution of $\mathbf{y}$ may be approximated by the multi-variate Gaussian distribution.

Note that the (i,j)-entry of the $l \times l$ covariance matrix of $\mathbf{y}$, R, is given by

$$R(i,j) = (1+p)^{j-i}\frac{\sigma_{k+i}}{\sigma_{k+j}} + \frac{\sigma_{\mathbf{w}}^2}{\sigma_{k+i}\sigma_{k+j}}\delta_{i-j}.$$

Now that we computed the covariance matrix R, the probability density function of $\mathbf{y}$ can be approximated by the multi-variate Gaussian distribution

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{(2\pi)^{l/2}(\det R)^{1/2}} e^{-\frac{1}{2}\mathbf{y}^T R^{-1}\mathbf{y}}. \quad (4)$$

The joint maximum-likelihood estimate of x_0 and p can be found by solving the maximization problem

$$\min_{x_0,p}\{\mathbf{y}^T R^{-1}\mathbf{y} + \log\det R\}. \quad (5)$$

On the other hand, the traditional approach to the estimation of the initial population in a branching process first focuses on finding the maximum-likelihood estimator of p [8],

$$\hat{p} = \frac{z_{k+1}+\cdots+z_{k+l}}{z_k+\cdots+z_{k+l-1}} - 1. \quad (6)$$

Then, the above estimate $\hat{p}$ is used to estimate x_0 as

$$\hat{x}_0 = \frac{z_{k+l}}{(1+\hat{p})^{k+l}}. \quad (7)$$

Note that for the reliability of the estimate $\hat{p}$ in (6), we only used measurements taken in the exponential phase of RT-PCR. Also, note that the objective function of the optimization (5) is not convex. To solve it, one can use, e.g., a gradient search initialized by $\hat{x}_0$ and $\hat{p}$ obtained from (7) and (6), respectively.

In Figure 1, we compare the mean-square error of the estimate of x_0 computed by (5) and that of (7) for the case of two measurements in the exponential phase ($l = 2$). We see that for the particular set of parameters ($x_0 = 100$, the noise variance 1/100 of the signal intensity), the joint maximum-likelihood estimator (5) outperforms the estimator (7) over the considered range of values of p by an order of magnitude.

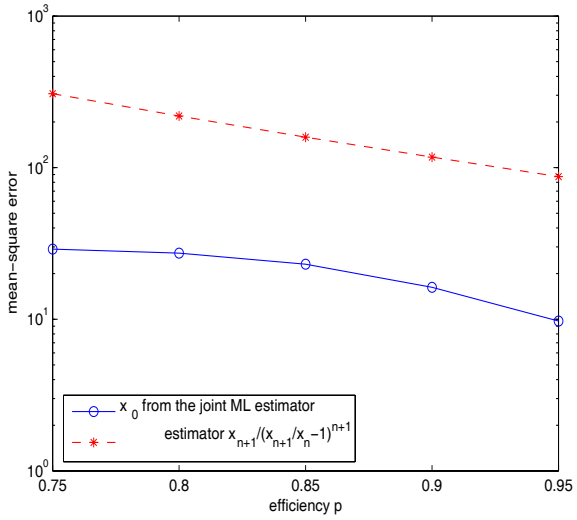

Fig. 1. *Comparison of the estimation mean-square errors*

2. REFERENCES

[1] K. Mullis and F. Faloona, "Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction," *Methods Enzymol* 155:335-350, 1987.

[2] M. A. Innis, D. H. Gelfand, and J. J. Sninsky, *PCR Applications: Protocols for Functional Genomics*, Academic Press, 1999.

[3] G. Stolovitzky and G. Cecchi, "Efficiency of DNA replication in the polymerase chain reaction," *PNAS*, vol. 93, pp. 12947-12952, November 1996.

[4] C. Jacob and J. Peccoud, "Estimation of the parameters of a branching process from migrating binomial observation," *Adv. in Applied Prob.,* 30,948-967, 1998.

[5] N. Lalam, C. Jacob, and P. Jagers, "Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency," *Adv. in Applied Probability*, 36, 602-615, 2004.

[6] A. Hassibi, H. Kakavand, and T. H. Lee, "A stochastic model and simulation algorithm for polymerase chain reaction (PCR) systems," GENSIPS 2004.

[7] D. Wang et. al., "Estimating the mutation rate during error-prone polymerase chain reaction," *J. of Comput. Biology*, 7, 143-158, 2000.

[8] J.-P. Dion, "Estimation of the mean and the initial probabilities of a branching process," *Journal of Applied Probability*, 11, 687-694, 1974.

[9] H. Cramer, *Mathematical Models of Statistics*, Princeton University Press, Princeton, NJ 1946.